Determining factors for health insurance charge

ULLAS UMESH

Table of Contents

Introduction

In this study, we aim to unravel the intricate relationships between insurance charges and an array of variables, providing a deeper understanding of the factors that contribute to the cost of medical coverage. The primary goal of this analysis is to dissect the myriad factors influencing insurance charges. By examining personal attributes, lifestyle choices, and regional disparities, we seek to uncover patterns that can inform decision-making in healthcare planning, policy formulation, and individual financial planning.

Our analysis leverages a robust dataset encompassing a diverse group of individuals. The dataset includes variables such as age, gender, BMI (Body Mass Index), number of children, smoking habits, geographic region, and the pivotal variable – medical insurance charges incurred by everyone (Narayana and Kowshik).

Variable conditions that affect the medical charges:

1. Age: Investigating how age influences insurance charges, recognizing the potential impact of healthcare needs across different life stages (Blundell, et al. 2024)

2. Gender: Analysing gender-related disparities in charges, acknowledging potential variations in healthcare utilization and costs.

3. BMI: Exploring the intricate relationship between body mass index and medical expenses, considering implications for preventive care and chronic conditions (Ward, et al. 2021).

4. Children: Assessing the financial implications of family size, as the number of dependents may influence healthcare needs and costs.

5. Smoking Habits: Investigating the correlation between smoking habits and insurance charges, recognizing the health risks associated with smoking (Anon. 2022).

6. Geographic Region: Understanding regional variations in insurance costs, considering differences in healthcare infrastructure, cost of living, and regional health patterns.

Our analysis employs a robust methodology, incorporating statistical techniques, data visualization, and hypothesis testing. By utilizing tools such as descriptive statistics, regression analysis, and interactive visualizations, we aim to present a nuanced and comprehensive perspective on the factors affecting insurance charges. Understanding the factors that drive insurance charges is crucial for stakeholders across the healthcare landscape. “According to various studies, major factors that contribute to higher expenses in personal medical care include smoking, aging and BMI” (Alfons and Francis 2020). This analysis strives to contribute valuable insights to policymakers, insurers, healthcare providers, and individuals, fostering informed decision-making and promoting a more equitable and sustainable healthcare system.

Cleaning the dataset

In the realm of data analysis, the significance of data cleaning cannot be overstated. As a pivotal stage in the data preprocessing pipeline, it ensures that the dataset is refined, accurate, and primed for insightful analysis. To embark on the data cleaning journey, a thorough understanding of the dataset is paramount. This entails familiarizing oneself with variable names, types, and their contextual relevance. A deeper comprehension of each variable's meaning sets the stage for informed decision-making during the cleaning process (Swapna, et al.).

The first hurdle to overcome is the presence of missing values. Utilizing summary statistics and visualization techniques, the data analyst pinpoints the locations and extent of missingness. This step serves as a foundation for subsequent decisions regarding imputation or removal of missing values. A thorough check for missing values was carried out using is.na( ) method and we found that there were no missing values in the dataset (Langkamp, Lehman and Lemeshow). Next, we checked for duplicate records and how they can distort analysis results. Identifying and removing duplicates, ensures that each observation contributes distinct information to the analysis. Outliers, while potentially valuable, can skew

results (Mahapatra, et al. 2020). Deciding whether to remove outliers or apply transformation methods ensures that extreme values do not unduly influence subsequent analyses. Standardizing or normalizing numerical variables brings uniformity to the data, facilitating fair comparisons between variables with different scales. Techniques like Min-Max scaling or Z-score normalization are common in this step. Ensuring that variables possess the correct data types is crucial for subsequent analyses. Converting strings to categories or dates to datetime objects enhances the dataset's suitability for analysis. Categorical variables require special attention. One-hot encoding or label encoding ensures that these variables are appropriately represented, allowing for meaningful statistical analyses. Inconsistencies in data entry can introduce errors. Identifying and rectifying inconsistencies, guided by domain knowledge, promotes accuracy and reliability.

Upon satisfaction with the cleaning process, the finalized dataset is saved for further analysis. Organized and ready for use, the cleaned data becomes a reliable foundation for drawing meaningful insights. In conclusion, the data cleaning process is a meticulous journey that transforms raw data into a refined and reliable asset.

Descriptive Analysis

Table 1: Summary Statistics of all Continuous Variables

Variables	Min.	Max.	1st Quartile	3rd Quartile	Mean	Median	Standard deviation	Variance	Kurtosis
Age	18.00	64.00	27.00	51.00	39.21	39.00	14.0499	197.40	1.7550
BMI	15.96	53.13	26.30	34.69	30.66	30.40	6.0981	37.18	2.9449
Children	0.0	5.0	0.0	2.0	1.095	1.00	1.2054	1.45	3.1972
Charges	1122	63770	4740	16640	13270	9382	12110	146652372	4.5958

The age distribution ranges from 18 to 64 years, with a central tendency around 39 years. As per the graphs shown in APPENDIX I, the relatively small kurtosis suggests a moderately peaked distribution, indicating that age values are somewhat concentrated around the mean. The BMI values range from 15.96 to 53.13, with a central tendency around 30.66. A positive kurtosis score suggests that the distribution of BMI data has heavier tails, which indicates the presence of probable outliers or extreme results. Since there is a dependent factor i.e., smoking which impacts on the physical health of an individual and in turn affects the BMI, we shall ignore the outliers. The number of children covered in the insurance ranges from 0 to 5, with a central tendency around 1.095. The positive kurtosis indicates a distribution with heavy tails, suggesting potential outliers or extreme values in the number of children data. The insurance charges vary from 1122 to 63770, with a central tendency around 13270. The positive kurtosis indicates a distribution with heavy tails, suggesting potential outliers or extreme values in the insurance charges data. The positive kurtosis in all the variables indicates potential outliers.

Table 2: Frequency count of Categorical Variable

Variables	Categories	Frequency
Sex	male	676
	female	662
Smoker	yes	274
	no	1064
Region	northeast	324
	northwest	325
	southeast	364
	southwest	325

The frequency distribution for sex indicates a balanced representation of gender in the dataset, with a slight predominance of males. The Smoker variable indicates a notable imbalance, with a higher frequency of non-smokers in the dataset. The Region variable suggests a somewhat balanced representation of individuals across different geographic regions.

Statement of Problem

The statistical analysis encompasses several key steps, starting with exploring dependencies and evaluating the strength of relationships among the predictor variables to prove that they are independent. The next phase involves fitting a linear regression model, allowing for a comparison with multiple linear regression. Through standard statistical procedures, we thoroughly examine our findings and results. Subsequently, we assess the differences in central tendencies among independent variables and newly formed 'CHARGE_split' variable. Finally, using ANOVA and Kruskal-Wallis Testing we find out the differences in predictor variable with respect to geography. This systematic approach enables us to draw valid conclusions based on the observed patterns and relationships within the dataset.

Statistical Analysis

As stated in the problem statement, the first step of analysis would be to prove that the predictor variables are independent of each other. We first evaluate the type of test to be performed on each variable by performing the test for Normality (Das, Rahmatullah Imon and Imon 2016).

Normality test

The hypotheses for the tests are:
Null hypothesis(H_0): Each predictor variable is from a population that is normally distributed.
Alternative hypothesis(H_1): Each predictor variable is **not** from a population that is normally distributed.
The threshold for decision making is 5% level of significance. **α = 0.05**
Kolmogorov-Smirnov test is carried out using,

$$D = \max\left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)\right)$$

$1 \le I \le N$

The p value is calculated using the above test statistic.

Table 3: Result for Normality Test

Predictor Variables	p-values	Decision	Type of test
Age	1.114×10^{-7}	Given that the p-value is less than 0.05, we reject H_0 and conclude that the sample came from a population without a distribution.	Non-Parametric test
BMI	0.3313	Given that the p-value is greater than 0.05, we don't have enough evidence to reject the H_0 and conclude that the sample came from a normally distributed population.	Parametric test
Children	2.2×10^{-16}	Given that the p-value is less than 0.05, we reject H_0 and conclude that the sample came from a population without a distribution.	Non-Parametric test
Charges	2.2×10^{-16}	Given that the p-value is less than 0.05, we reject H_0 and conclude that the sample came from a population without a distribution.	Non-Parametric test

The null hypothesis was rejected for Age, Children and Charges since their respective p-values were less than the threshold and we conclude that we have enough evidence to say that the sample were drawn from a distribution free population. On the other hand, the p-value stood at 0.3313 for BMI which clearly shows that the population is normally distributed. Furthermore, APPENDIX I shows the bell-shaped curve for BMI vs density plot which further reinforces our null hypothesis.
The choice of test type:
1) If the variable failed normality test, a Non-Parametric test will be carried out.
2) If the variable passed normality test, a Parametric test will be carried out.

Test for Correlation

Null hypothesis(H_0): there is no correlation between any two variables.

Alternative hypothesis(H_1): linear association exist between any two variables.

The threshold is 5% level of significance.

The test statistics are Pearson and Spearman for parametric and non-parametric respectively.

Pearson test,

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Spearman test,

$$\rho = 1 - \frac{6 \sum {d_i}^2}{n(n^2 - 1)}$$

Table 4: Test for various Correlations and Decisions

Continuous Variables	Correlation Coefficient	Type of Test	p-values	Decisions
Age & BMI	0.1092	Pearson	6.194×10^{-5}	Based on the p-value being less than 0.05, we reject the null hypothesis H_0 and conclude that there is a statistically significant connection between the two variables.
		Spearman	7.859×10^{-5}	
Age & Children	0.0424	Spearman	0.03712	Based on the p-value being less than 0.05, we reject the null hypothesis H_0 and conclude that there is a statistically significant connection between the two variables.
Age & Charges	0.2990	Spearman	2.2×10^{-16}	Based on the p-value being less than 0.05, we reject the null hypothesis H_0 and conclude that there is a statistically significant connection between the two variables.
BMI & Children	0.0127	Pearson	0.641	Since p-value is greater than 0.05, we don't have enough evidence to reject H_0 and conclude that there is no linear relationship between both variables.
		Spearman	0.5684	
BMI & Charges	0.1983	Pearson	2.459×10^{-13}	Based on the p-value being less than 0.05, we reject the null hypothesis H_0 and conclude that there is a statistically significant connection between the two variables.
		Spearman	1.193×10^{-5}	
Children & Charges	0.0679	Spearman	9.847×10^{-7}	Based on the p-value being less than 0.05, we reject the null hypothesis H_0 and conclude that there is a statistically significant connection between the two variables.

Except for BMI and Children all other variables have correlation between both variables as they possess p-value < 0.05, which is statistically significant. We also got an idea that there is a strong relation between Age & Charges as well as BMI & Charges as their p-values were significantly lower compared to other variable pairs. Refer APPENDIX III for visual representation of the correlation.

Chi-Square Test was carried out to find out correlations between the categorical variables.

Table 5: Chi – Square Test for Categorical variables

Categorical Variables	Type of Test	p-value	Decisions
Sex & Smoker	Chi – Square	0.006548	Given that the p-value is below 0.05, we may reject the H_0 and state that there is a statistically significant correlation between the two variables.
Sex & Region	Chi – Square	0.9329	Given that the p-value is greater than 0.05, we do not have enough evidence to reject the H_0 and state that there is no statistically significant connection between the two variables.
Smoker & Region	Chi – Square	0.06172	Given that the p-value is greater than 0.05, we do not have enough evidence to reject the H_0 and state that there is no statistically significant connection between the two variables.

After performing Chi-Square test, we found that there was a slight correlation between the Sex and Smoker categorical variables.

Regression Analysis

Further Multiple linear Regression was carried out to examine the individual and collective contribution of the predictor variable to the variation in the dependent variable.

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n$$

Table 6: Multiple Linear Regression results

Coefficients	Estimate Std.	Error	t value	Pr (>| t |)
Age	256.86	11.90	21.587	$< 2 \times 10^{-16}$
Sex_male	-131.31	332.95	-0.394	0.693348
BMI	339.19	28.60	11.860	$< 2 \times 10^{-16}$
Children	475.50	137.80	3.451	0.000577
Smoker_yes	23848.53	413.15	57.723	$< 2 \times 10^{-16}$
Region_northeast	960.05	477.93	2.009	0.044765
Region_northwest	607.09	477.20	1.272	0.203533
Region_southwest	-74.97	470.64	-0.159	0.873460

As visualised in APPENDIX IV, age has a positive coefficient, indicating that, on average, as age increases by one unit, medical insurance charges are expected to increase by 256.86 units. The low p – value suggests that the ag e variable is statistically significant in predicting charges.

The coefficient "Sex_male" represents effect of being male. The negative coefficient suggests that being male is associated with a decrease of 131.31 units in medical insurance charges. However, the hight p – value suggests t hat this variable is not statistically significant in predicting charges.

BMI has a positive coefficient, suggesting that, on average, as BMI increases by one unit, medical insurance char ges are expected to increase by 339.19 units. The low p-value indicates that BMI is statistically significant.

An additional child is associated with an increase of 475.50 units in insurance charges but, the low p-value sugge sts that the variable is less significant. The large positive coefficient indicates that being a smoker is associated with a substantial increase in insurance charges. The low p -value not only indicates the statistical significance b ut also shows that the smoker factor increases the insurance charges exponentially. The low p -value for the nort heast region signifies the slight statistical significance.

Test for Differences

At first, we created a new categorical variable naming CHARGE_split which categorizes the charges variable into two categories namely, 'Category A' and 'Category B'. These categories were distinguished based on the median value of the charges variable. Later, we move on to find the differences in central tendencies of these two categories. We perform Normality test to determine the type of test to be carried out based on the nature of distribution.

Table 7: Test for Normality for various Categories

Continuous Variables	Category	p-value	Type of Test
Age	Category A	8.191×10^{-5}	Non - parametric
	Category B	7.438×10^{-14}	
BMI	Category A	0.5191	Parametric
	Category B	0.2206	
Children	Category A	p-value $< 2.2 \times 10^{-16}$	Non - parametric
	Category B	p-value $< 2.2 \times 10^{-16}$	
Charges	Category A	0.0002969	Non - parametric
	Category B	p-value $< 2.2 \times 10^{-16}$	

After determining the type of test to be carried out, we decided to proceed with Mann Whitney U test for Non – parametric variables and t-test for Parametric variables (Kresojević and Gajić 2019).
Null hypothesis(H_0): there is no difference between the distribution of the two categories.
Alternative hypothesis(H_1): there is difference between the distribution of the two categories.
The threshold is 5% level of significance.

Mann Whitney U Test,

$$U_i = n_1 n_2 + \frac{n_i(n_i + 1)}{2} - R_i$$

t Test,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Table 8: Test for Differences

Independent Variable	Type of test	p-value	Decision
Age	Mann Whitney U test	$< 2.2 \times 10^{-16}$	Given that the p-value is below 0.05, we may reject the H_0 and state that statistical difference exists between the population mean of both groups.
BMI	t - test	0.0009952	Given that the p-value is below 0.05, we may reject the H_0 and state that statistical difference exists between the population mean of both groups.
Children	Mann Whitney U test	0.7154	Given that the p-value is greater than 0.05, we do not have evidence to reject the H_0 and state that there is no difference between the population median of both groups.
Charges	Mann Whitney U test	$< 2.2 \times 10^{-16}$	Given that the p-value is below 0.05, we may reject the H_0 and state that statistical difference exists between the population mean of both groups.

After running the test for differences, we found that there were significant changes in central tendencies for both the categories for the independent variables Age, BMI and Charges where, there is no difference for the variable Children.

Test for Central Tendencies based on Geographic area.

As stated in the problem statement the final step of the analysis was to find the differences in central tendencies based on the geographic region. We carry out this analysis using Kruskal-Wallis for Non-parametric variable and ANOVA test for parametric variable.
Kruskal – Wallis Test,

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

ANOVA test,

$$F = \frac{MS_{between}}{MS_{within}}$$

Table 9: Test to determine differences in central tendencies with respect to region.

Independent Variable	Type of Test	p-value	Decision
Age & Region	Kruskal - Wallis	0.9374	Given that the p-value is greater than 0.05, we do not have enough evidence to reject the H_0 and state that there is no significant difference in age distribution among the regions.
BMI &Region	ANOVA	$< 2 \times 10^{-16}$	Given that the p-value is below 0.05, we may reject the H_0 and state that that there is significant difference in BMI among the regions.
Children & Region	Kruskal - Wallis	0.4982	Given that the p-value is greater than 0.05, we do not have enough evidence to reject the H_0 and state that there is no significant difference in children distribution among the regions.

As per ANOVA test carried out on BMI & Region and after plotting the graphs represented in APPENDIX VI, we understood that there is significant difference in BMI with respect to different regions. Further we carry out post hoc test to find out the range within which the difference exists between the 4 regions.

Table 10: Post hoc test results

Region	Diff	Lwr	Upr	P adj
Northwest-northeast	0.02628153	-1.1552239	1.207787	0.9999328
Southwest-northeast	4.18248592	3.0330135	5.331958	0.0000000
Southwest-northeast	1.42311230	0.2416069	2.604618	0.0106965
Southeast-northwest	4.15620440	3.0076679	5.304741	0.0000000
Southwest-northwest	1.39683077	0.2162360	2.577426	0.0127393
Southwest-southeast	-2.75937363	-3.9079101	-1.610837	0.0000000

Post hoc test helped us determine the pair of regions with statistical significance difference in mean BMI. The adjusted p-value for southwest-northwest and southwest-northwest indicates that the statistical significance in mean BMI of these pairs while, Northwest-northeast has no significance in mean BMI.

Conclusion

Our examination of health insurance data has illuminated the substantial impact of age, BMI, smoking status, and region on insurance costs. Age and BMI emerged as pivotal determinants, exhibiting a direct correlation with increased insurance charges. Notably, smoking status markedly elevates costs, underscoring the profound influence of lifestyle choices on financial burdens. Moreover, our regional analysis unveiled significant variations in BMI, indicative of regional lifestyle disparities influencing health outcomes. Rigorous tests for variable independence and meticulous statistical analyses bolster the reliability of our findings. The discernible discrepancies identified, particularly in BMI across regions, underscore the imperative for tailored health policies in different locales. In summary, our study furnishes valuable insights into the drivers of health insurance charges, facilitating informed decision-making for insurers, policyholders, and policymakers alike. Comprehensive details and illustrative graphics are provided in the appendices, enhancing accessibility and comprehension of the intricate dataset (Kaushik, et al. 2022).

References

Adult smoking habits in the UK methodology Methodology information for the Adult smoking habits in the UK annual statistical bulletin. 2022.

ALFONS, A. and FRANCIS, K., 2020. *Linear regression model for predicting medical expenses based on insurance data*.

BLUNDELL, R., et al., 2024. *Old Age Risks, Consumption, and Insurance*. American Economic Association.

DAS, K.R., RAHMATULLAH IMON, A.H.M. and IMON, A.H.M.R., 2016. *A Brief Review of Tests for Normality*. Science Publishing Group.

KAUSHIK, K., et al., 2022. *Machine Learning-Based Regression Framework to Predict Health Insurance Premiums*. MDPI AG.

KRESOJEVIĆ, B. and GAJIĆ, M., 2019. *Application of the T-Test in Health Insurance Cost Analysis: Large Data Sets*. Walter de Gruyter GmbH.

LANGKAMP, D.L., LEHMAN, A. and LEMESHOW, S., *Techniques for Handling Missing Data in Secondary Analyses of Large Surveys*.

MAHAPATRA, A.P.K., et al., 2020. *Concept of Outlier Study: The Management of Outlier Handling with Significance in Inclusive Education Setting*. Sciencedomain International.

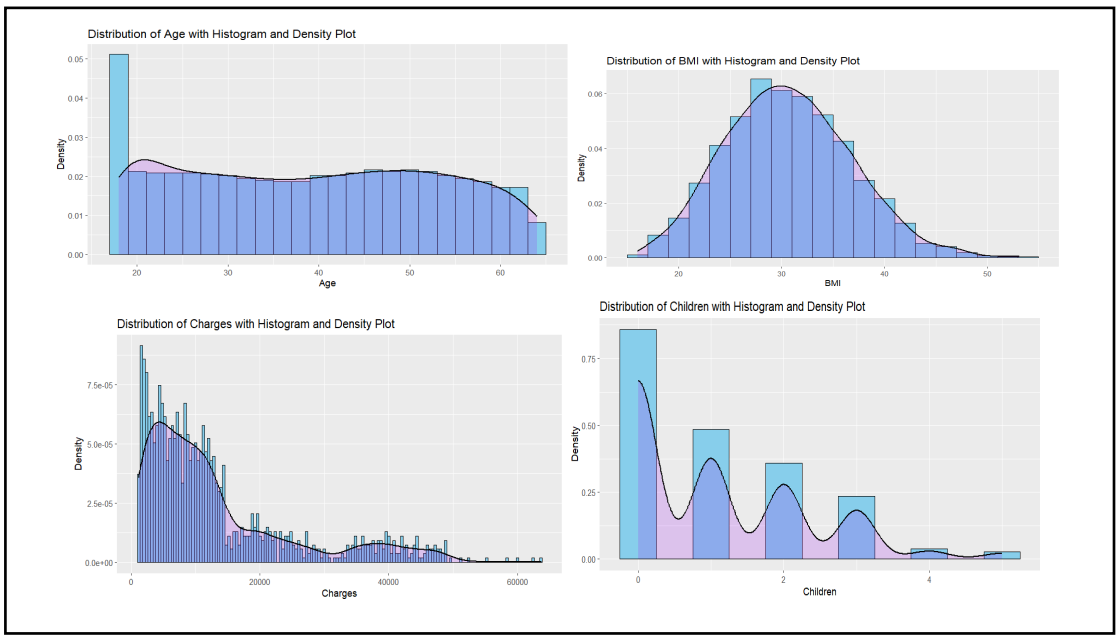
NARAYANA, K.L. and KOWSHIK, P., *Medical Insurance Premium Prediction Using Regression Models*.

SWAPNA, S., et al., *Data Cleaning for Data Quality*.

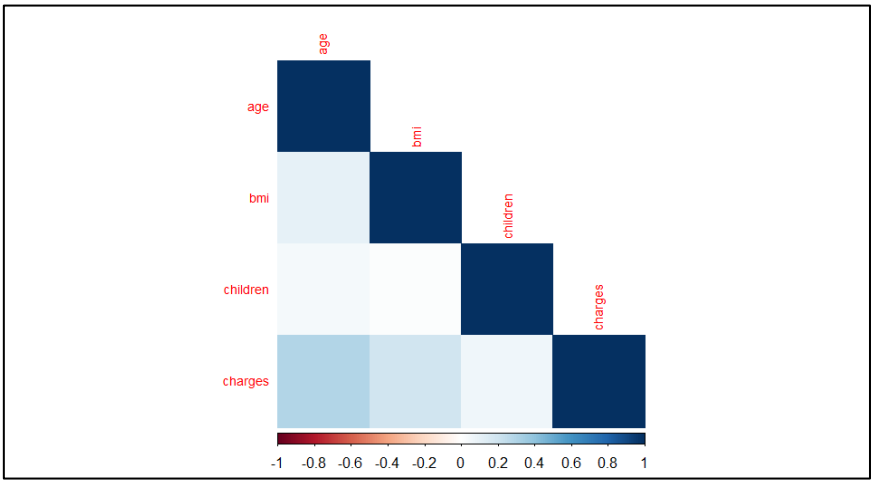
WARD, Z.J., et al., 2021. *Association of body mass index with health care expenditures in the United States by age and sex*. Public Library of Science (PLoS).

APPENDICES

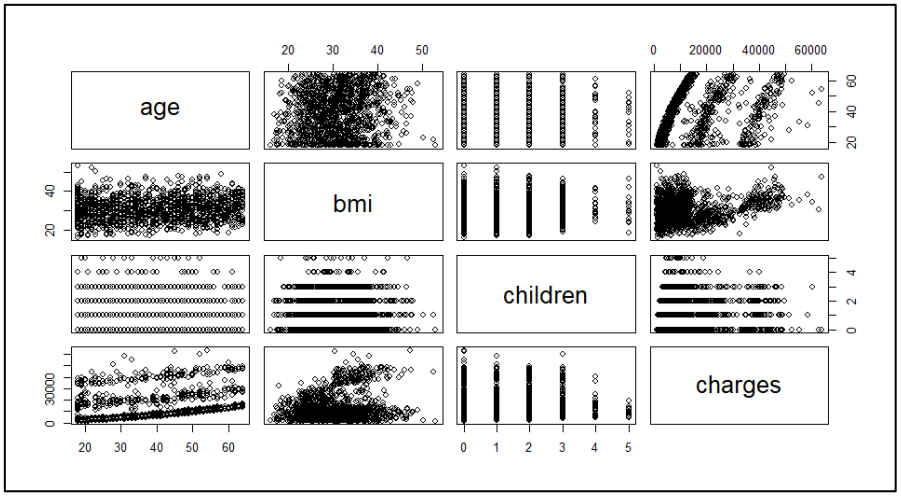
APPENDIX I – Histogram and Density plot of Continuous Variables



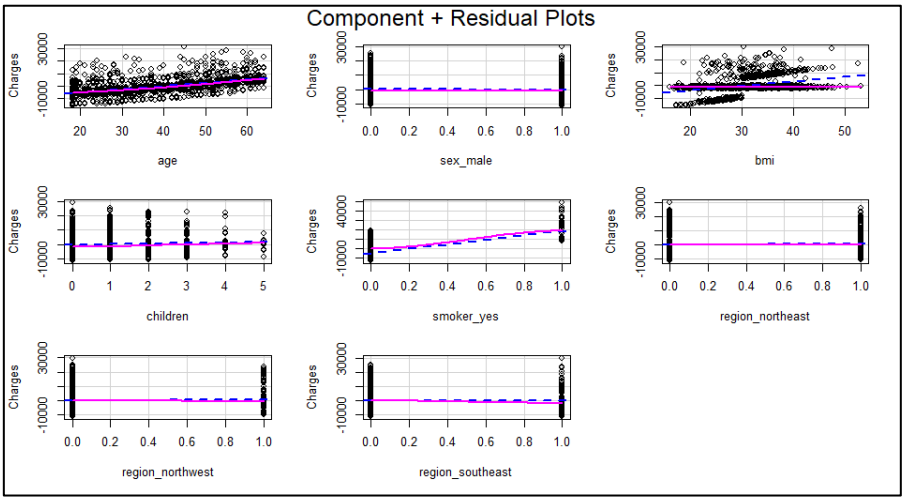
APPENDIX II



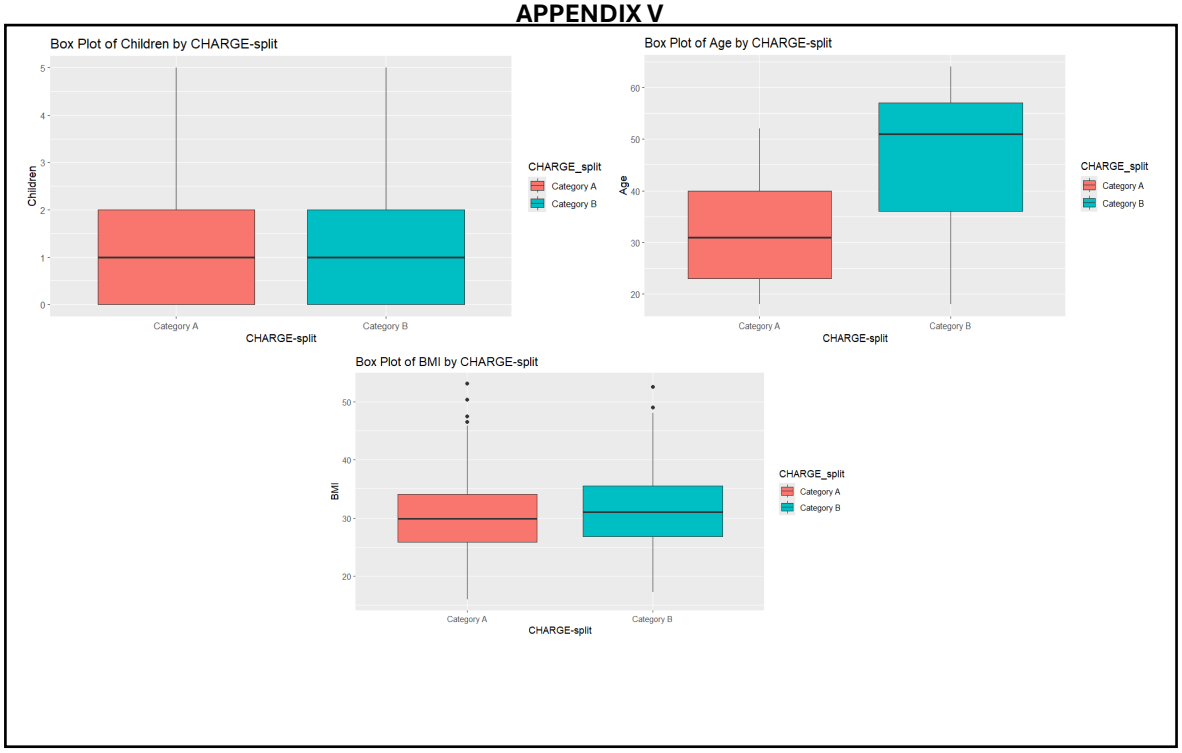
APPENDIX III – Scatter plot of Continuous Variables



APPENDIX IV – Component + Residual plot



APPENDIX V



APPENDIX VI